Argumentative Task Design

Target users

Taking the example by Sedova (2017) as the basis for our task, we set our target audience as high-school students, for three reasons, detailed below. Firstly, studies on the use of dialogue systems for education focus mostly on higher education instead of the lower levels (Kuyven et al., 2018).

Secondly, oracy skills are very important for students' academic and work life (Mercer et al., 2017). In most countries, the duration of compulsory education corresponds to the years needed to complete secondary education 910, which means that, after high school, some students may enter the workforce and not complete any more years of education, while others may pursue some form of higher education. In both cases, oracy skills would be necessary (Mercer et al., 2017), especially when students join a competitive, globalized academic world or labor market (Okada et al., 2018). Thus, oracy skills need to be developed in the lower stages of education to benefit all students. Thirdly, designing a system aimed at a younger audience (i.e. children instead of teenagers) would involve greater challenges that might be unnecessary in a preliminary study like this. The main reason for this is that speech recognition performs more poorly with children speech (Yu et al.); this is partly due to the distinct characteristics of children's speech (ibid), which could also complicate the task of designing responses of the appropriate level.

We are also designing the task specifically with American English-speaking students in mind due to the higher availability of resources for that setting, though the task might be suitable for English speakers from other countries or non-native English speakers.

Reference text selection

Linking the target audience with the example tasks described by Sedova (2017) brings to mind college admission tests, many of which involve analyzing and/or developing an argument { some examples are the American SAT (CollegeBoard, 2015), the German Abitur11 or the Spanish PAU12. As we have selected American high-school students for our target audience, we use the SAT as our basis for the task - as in the task described by Sedova (2017), a text is used to start a discussion. Below we further justify why we selected the SAT as our source for texts.

Texts as domain knowledge source. A task that can be performed by a wide variety of students may require a given text as its base, so that students with different background knowledge, exposed to different curricula and teaching methods, can have the same information needed to complete the task. Otherwise, if students using a dialogue system are just asked to discuss a topic, they may not know enough about the topic if it was not mentioned in their particular class. Therefore, the system will need to provide information. For example, Arguebot (Kulatska, 2019), one of the few existing dialogue systems that are similar to this project, provided a text with information on the topic of debate, even though the users could choose the topic; this information was considered useful by over 80% of testers (ibid). Those who did not find the text useful claimed that they would have preferred the information in a more streamlined format, such as an outline (ibid); however, it must be

borne in mind that Arguebot was not designed for pedagogical purposes, and thus the improvements it might require may not apply to dialogue systems with a different purpose.

Texts as models. Providing a text as basis for the argumentative task also gives students a model to learn from. According to Observational Learning Theory, analyzing a model can be conducive to learning; this theory is considered relevant especially with skills related to language, as these require the student to practice, because those skills cannot be learned through mere observation (unlike what its name may suggest, Observational Learning Theory involves more than observation, observation is only the first step) (Okada et al., 2018). Thus, the student needs to first observe the model, but then also analyze it, imitate it, and finally reflect on their performance, which could be fostered by being given feedback (ibid). In the next section we describe our proposed task: the first subtask being an introduction, subtask 2 would cover the observation of the model, subtasks 3 and 4 the scaffolded analysis, subtask 5 the imitation of the model, and feedback would be provided throughout the entire task.

Texts to reinforce dialogic principles and oracy skills. Including a text in the task also adds a voice aside from the student's and the dialogue system's. Ideally, a discussion should involve more than two participants (Alexander, 2010), but not enough progress has been made on dialogue systems that can interact with more than one user at the same time, distinguishing each person's contributions (Marge et al., 2020). However, even if the student is performing the task alone with the dialogue system, the inclusion of an author's voice contributes to some extent to the principles of reciprocity and collectiveness, which might otherwise be absent in a context of student-computer interaction. Additionally, this other voice could help students develop certain oracy subskills (Mercer et al., 2017) that might not be reinforced in a different kind of student-computer interaction. These oracy subskills are primarily the social subskills; a discussion where the text's author cannot answer limits the social aspect of dialogue, but it at least involves the student paying attention to someone else's views. Analyzing someone else's argumentative text could also reinforce the cognitive oracy subskills of evaluating and summarizing ideas and building on others' views. Also, as has been mentioned, a text can serve as a model, and this model could help the student enrich their pool of linguistic resources to strengthen their linguistic oracy subskills.

Texts from appropriate sources. Using the SAT texts as the basis for the task ensures that the texts are suitable for high-school students, challenging but not exceedingly difficult, and that the students can work on them without prior knowledge of the specific topic being discussed (CollegeBoard, 2015). SAT Essay texts are selected by CollegeBoard researchers so that they \argue a point", are \written for a broad audience" and \use logical reasoning and evidence to support claims", which makes them not only suitable for highschool students, but also specifically for the task of analyzing someone's argumentation13. The SAT exam is taken by millions of students { for example, in 2020 more than 2.1 million students took the SAT Essay test in the US (CollegeBoard, 2020); due to this large number of test takers, there is an abundance of resources which can be used to train the dialogue system, even if we only consider resources from the redesigned test { the redesigned version of the exam was introduced in 2016 to better align the test with students' and colleges' needs (CollegeBoard, 2015). The SAT Essay test does not require students to provide any arguments of their own,

as the focus is on seeing the students' reading and analytical skills as well as general writing skills (ibid). For our task to follow the principles of dialogic teaching and help students incorporate ideas into their own understanding and develop their own arguments, we must thus design a task that is different from the SAT Essay task, even if it takes the same texts as its basis. This difference between the SAT Essay test and our own task, however, means that the available sample answers will not be useful to us, which is one of the reasons why we are developing a dataset of artificial answers.

Task description

As explained above, we took the SAT as our source for texts to start an argumentative task; however, we needed to modify the SAT task to align it with dialogic teaching and so that it could be carried out with a dialogue system of the characteristics described in our framework. The concrete realization of the task that we arrived at, divided into five subtasks (ST), is as follows:

1. **Introduction and guidance**. Firstly, the dialogue system introduces itself and provides the student some guidance on how to use it and perform the task. We consider this guidance necessary for the student to be fully aware of the system's capabilities and thus use it to its full potential (Jokinen, 2009; Thies et al., 2017; Huang et al., 2019), as well as to ensure that the task is performed in the way most conductive to learning (Pinkwart et al., 2008). The system could also mention why the task might be valuable to the students to increase their motivation (Mercer and Howe, 2012; Thies et al., 2017).

2. **Text reading**. The actual task begins with the student reading the assigned argumentative text. The student is asked to read it out loud at least once for a voice recognition system to confirm that the student completed the reading; we believe this to be necessary because this type of task cannot succeed if the student does not complete the reading (Sedova, 2017), and students with little motivation to complete a task may only put in the minimum effort unless they are encouraged to perform the task in a more productive way (Pinkwart et al., 2008).

3. **Paragraph identification (thesis and arguments)**. Once the student has read the text, they are asked to identify the paragraph that contains the author's thesis; then, the main argument (or any argument if they are all given the same weight by the author). The identification is done by selecting the paragraph number, in order to avoid using Natural Language Processing where it serves no other purpose than complicating the system (Jokinen, 2009).

4. **Thesis and argument rephrasing**. The student is then told to put the information they just identified (thesis and one argument) into their own words to understand and evaluate the ideas (Skidmore, in Mercer at al., 2019). The system identifies whether the student's input is an actual attempt at an answer and, if so, it evaluates the answer on the feedback aspects we are considering (being on task, completeness, paraphrasing, correctness) and returns feedback that can guide the student towards a better answer. This is the subtask (ST4) we focus on for our dataset compilation and our experiment.

5. **Scaffolded discussion**. The system then asks the student to develop their own argument on the issue of the text. To make this a more scaffolded task, the system guides the student through several steps (choosing a stance, providing arguments to back it, responding to

counterarguments) until the student develops a solid argumentation (to determine when this goal is reached, quantity or quality criteria would need to be established). As tests concerning this part of the system are beyond the scope of this project, this part of the system is presented only as an option whose feasibility and suitability are to be confirmed in future studies.

- In the first step, the system would ask the student what their stance is on the topic under discussion. If the student were unsure, the system could provide links to sites such as Idebate, which give a quick overview of important arguments for and against a debate topic. At any point in the following parts of the task, the student could be allowed to signal that they have changed their mind, and their argumentation would restart from this first step.

- The system would then ask the student to provide an argument to defend their position. Again, sites like Idebate could be presented to students who cannot think of arguments to support that position; if the student defended the same position as the author of the model text, the system could also point the student to paragraphs where arguments were identified.

- The system would then search its knowledge base for arguments that could challenge what the student said. This knowledge base would need to contain a large number of arguments classified by topic and stance, and the aspect would also need to be identified to match it with the aspect of the student's argument (Gurevych, 2021). Some of IBM Debater's datasets, such as XArgMining14, could prove useful due to their size and topic/subtopic/stance/quality labels, if they could be aligned with the topics covered in the task.

    - If the system found no counterargument, it could simply ask the student to provide an additional argument.

    - If the system found a counterargument, it would present it to the student and ask them to provide their rebuttal. If adequate data was available, the system could give the student resources to build their rebuttal in case they are unable on their own. In either case, after the student provided a rebuttal or failed to do so, the system could ask the student to continue providing arguments to strengthen their argumentation.

- The system could end this part of the task after a specific number of arguments and/or rebuttals were reached – the specific number would have to be decided based on the data available, the topic and the level of the students. If argument quality were measured, the task could be ended after a specific quality score were reached.

- The student would then be asked to place their arguments and rebuttals of the system's counterarguments in a diagram, as these tools have proven useful for students' argumentation (Pinkwart et al., 2008). The student could then use the

diagram to help them participate in class discussion, write an argumentative essay or simply as prove of task completion or as a way to reflect on their performance.

This task has been thought of as a scaffolded task, with a focus on guiding students to help them improve their skills. For that reason, feedback is to be given throughout the whole task as needed, and the student would need to be able to access resources to help with the task. Feedback thus takes a formative role, rather than summative. Nonetheless, tests with students and teachers would have to be carried out to ascertain whether additional feedback upon task completion might be useful.

System persona

It is necessary to design a personality for the system: users will inevitably project one as a consequence of the Media equation theory, by which people tend to treat computers as people (Google, 2021; Thies et al., 2017), or they may ask questions to learn about the system's personality instead of focusing on the task (Bii et al., 2013). If the persona is well designed, its effect on the user will be more easily controlled (Google, 2021). Due to the scope of this project, not much time can be devoted to this aspect. Tentatively, following the steps suggested by Google's guidelines and the insights from the literature on how young users interact with dialogue systems, the persona could be as described below, putting special emphasis on the general characteristics and not so much on the details such as the name, visual representation and specific voice type.

Name: Robosan (tentative gender-neutral name pending change after data on user preferences can be analyzed).

Characteristics: friendly, not trying too hard to be funny, non-judgmental but not overly supportive (Thies et al., 2017); empathetic (Goda et al., 2014; Thies et al., 2017) { using varied prompts can make the system sound less “robotic” and reinforce this illusion of empathy (Google, 2021).

Visual representation and voice: pending definition after data on user preferences can be analyzed

Interface

As can be seen in more detail in the example conversation, the suggested system interface is meant to contain more than only a chat window - the example only shows the chat, but other windows are referenced. As the task is based on a text, we suggest a separate window where the student could check the text more comfortably. We have previously mentioned that diagrams have been shown to be useful tools for argumentation (Pinkwart et al., 2008). For that reason, we suggest the inclusion of a diagram tool that the students can use to organize the ideas that they come up with while speaking with the dialogue system. Ideas on how to design this diagram tool could be taken from already existing tools like LARGO (ibid), which would need to be adapted to a younger and less specialized target audience.

In order to keep track of student progress, a user authentication component is also suggested. This might help students feel more accountable (Major et al., 2018) and it might also ease teachers' monitoring of student work.

We also propose including microphone controls in the interface, so the students can signal when they want to start their turn and be heard by the system. This proved to be a useful feature in the Wizard-of-Oz study carried out by Catania et al. (2020). The students in that study were children with no experience using dialogue system, which the authors attribute to these systems not being widely used in Italy, where the study was carried out. Our target students may be expected to have higher skills and familiarity with dialogue systems (ibid). Still, given that our proposed system is meant to scaffold students so that they can develop their oracy skills, microphone controls could also be a positive feature, as this could bring their attention to turn-taking, an aspect of the social and emotional oracy skills (Mercer et al., 2017).

Adaptations to increase accessibility, though encouraged for everyone's benefit (Jokinen, 2009), are beyond the scope of this project and we thus offer no suggestions in this respect.